LA-UR-97-1025
March 18, 1997

Nonparametric Variance-based Methods of Assessing Uncertainty Importance

by

Michael D. McKay
Los Alamos National Laboratory
Los Alamos, NM 87545–0600
mdm@lanl.gov

Appears in *Reliability Engineering and System Safety*, **57**, (1997) 267–279

Contents

List of Tables

List of Figures

Nonparametric variance-based methods of assessing uncertainty importance

by

Michael D. McKay

Los Alamos National Laboratory, Los Alamos, NM 87545–0600

Abstract

This paper examines feasibility and value of using nonparametric variance-based methods to supplement parametric regression methods for uncertainty analysis of computer models. It shows from theoretical considerations how usual linear regression methods are a particular case within the general framework of variance-based methods. Examples of strengths and weaknesses of the methods are demonstrated analytically and numerically in an example. The paper shows that relaxation of linearity assumptions in nonparametric variance-based methods comes at the cost of additional computer runs.

1 INTRODUCTION

Methods from statistical analysis are commonly used to quantify importance of input variables for computer models. Many types of analyses, for example, probabilistic risk assessments like those performed by the United States Nuclear Regulatory Commission,[1] rely on regression analysis methods which, for the most part, are predicated on assumptions of a linear analysis model. Although such analyses are often performed on rank-transformed data, the computations are those derived from linearity assumptions. The methods themselves, collectively called linear regression methods, form a cornerstone of statistical analysis. They have desirable robust properties relative to their assumptions. However, departures from linearity can cause serious degradation of the power of linear regression methods. Therefore, the validity of the assumptions of a linear analysis model, whether with raw or rank-transformed data, is of paramount importance to evaluating appropriateness of the analysis methodology.

Statistical analyses methods which rest on a set of conditions that involve minimal distributional assumptions are call nonparametric methods. In this paper, the analysis of the variance of a model output under the assumption of a parametric analysis model, namely, a linear model, is compared to a nonparametric analysis of the variance of the model output which is based on the properties of variance alone, and not on any particular relationship between model output and inputs. Nevertheless, some assumptions are needed, like the one that the variance exists. Whether (parametric) regression methods should be augmented or replaced by nonparametric methods depends on the situation at hand. Although this paper shows strengths and weaknesses of the methods, suggesting how they can be used to complement each other, the paper does not provide a definitive answer which would single out one method when results would be based on the usual situation of limited computing resources and sample sizes.

2 UNCERTAINTY AND IMPORTANCE

A model prediction y is the result of a computation within a model $m(\cdot)$ for a vector of input values x,

$$y = m(x)\,. \tag{1}$$

For simplicity of presentation, y is treated as a scalar quantity. The prediction uncertainty in y is determined by the triple

$$(f_x, V, m(\cdot)) \tag{2}$$

where the inputs x take on values in V with probability (density) function f_x. The prediction uncertainty in y is characterized by its induced probability (density) distribution f_y, called the prediction distribution. That is, prediction uncertainty is determined and characterized by

$$\begin{aligned} x &\sim f_x(x)\,,\ x \in V \\ y &= m(x) \\ y &\sim f_y(y)\,. \end{aligned} \tag{3}$$

From the formulation above, the entire vector x determines a unique value of y. (Stochastic simulation models require a different interpretation). Therefore, as a whole, the input vector x is completely important because fixing it at a single value reduces the (conditional) prediction distribution f_y to a single point. The more important question one asks is how important is a subset of input variables x^s, say, with regards to the prediction distribution. Using variance as a yardstick of importance, the question becomes, how much does the prediction variance decrease (on average) when the values of a subset of inputs x^s are held fixed.

The "variance importance" of inputs just discussed is a special case of ways to look at the difference between the prediction distribution f_y and the family of conditional prediction distributions $\left\{f_{y|x^s};\ x^s \in V_{x^s}\right\}$ determined by a subset x^s of input variables and indexed on their values. Because the (marginal) prediction distribution of y can be written as

$$f_y(y) = \int f_{y|x^s}(y) f_{x^s}(x^s) dx^s\,, \tag{4}$$

the importance of the subset x^s is related to the difference among the members $f_{y|x^s}(y)$ of the family of conditional distributions.

A common approach used in investigation of eqn 4 involves approximating the regression function $E(y \mid x^s)$, the mean of the conditional distribution of y on the right of the equation as a function of x^s. The functional dependence of y on a subset of input variables might be approximated using ordinary (stepwise) regression, often on the ranks of the variable values. The assumed form of the regression function is part of the issue of the analysis model which is discussed later in this paper. The approach has been applied by Iman and Hora[2] to the analysis of a fault tree model using a polynomial function for the conditional mean value of y. Saltelli, Andres, and Homma[3] review methods of the approach, and Homma and Saltelli[4] review and extend the nonparametric approach of Sobol'[5] and the best additive fit in Stein.[6]

With nonparametric variance-based methods, the (prediction) variance from the left hand side of eqn 4 is written in terms of the conditional variance from the right hand side, without any assumptions about the regression function and the functional relation between y and x^s. Various (conditional) prediction-variance ratios are used in variance-based methods by McKay[7] to provide measures of importance.

3 LINEAR REGRESSION METHODS

The following brief review of regression methods shows how they are a particular case of variance-based methods. Let the input (row) vector

$$x = (x_1, x_2, \cdots, x_p) \tag{5}$$

represent p inputs to a computer code. In a risk analysis, the output y might be a probability, like $y = y_u =$ Pr(Number of Cancer Deaths $> u$). If the computer code is represented by $m(\cdot)$, then we denote an actual code calculation or computation model by

$$y = m(x)\,. \tag{6}$$

The linear analysis model assumes that there is a (column) vector of unknown constants

$$\beta = (\beta_1, \beta_2, \cdots, \beta_p)^t \tag{7}$$

such that an approximation linear in x to the computation model $m(\cdot)$ is sufficient for the purposes of statistical analyses. That is,

$$\begin{aligned} y = x\beta &= \sum_{i=1}^{p} \beta_i x_i \\ &\simeq m(x)\,. \end{aligned} \tag{8}$$

It is to be understood that a constant term may be included in the model, for example, by introducing β_0 and $x_0 \equiv 1$.

For an arbitrary input (subset) x, the *linear analysis model* is

$$\begin{aligned} E(y \mid x) &= x\beta \\ y &= x\beta + e \\ E(e) &= 0 \\ Cov[e, x\beta] &= 0\,. \end{aligned} \tag{9}$$

The error term e in the linear analysis model of eqn 9 is usually treated as a random variable independent of x and having mean value zero. As applied in the analysis of computer models, the error term is actually the difference between the code calculation $m(x)$ and the linear approximation $x\beta$. The phrase "linearity assumption" denotes the analysis model in eqn 9, with allowance for the possibility that $Cov[e, x\beta] \neq 0$.

In the linear analysis model, the variance of the output y is expressible as a linear combination of the variances and covariances of the inputs x. First of all, the variance of y in eqn 9 is given by

$$V[y] = V[x\beta] + V[e] + 2Cov[x\beta, e]. \tag{10}$$

Under the (questionable) assumption that x and e are independent, the covariance term vanishes, leaving

$$V[y] = V[x\beta] + V[e]\,. \tag{11}$$

For independent components of x, this variance of y can be written as

$$V[y] = \sum_{i=1}^{p} \beta_i^2 V[x_i] + V[e] \,. \tag{12}$$

In general, though, not all of the inputs are independent. Therefore, the variance of y takes on a more complicated form which includes covariance terms,

$$V[y] = \sum_{i=1}^{p} \beta_i^2 V[x_i] + 2 \sum_{i=1}^{p} \sum_{j<i}^{p} \beta_i \beta_j Cov[x_i, x_j] + V[e] \,. \tag{13}$$

In general matrix notation, eqn 13 is written as

$$V[y] = \beta^t V[x] \beta + V[e] \,. \tag{14}$$

A reasonable measure of the importance of an input variable x_i that is independent of the other inputs is the term in eqn 12 that corresponds to it. That term is $\beta_i^2 V[x_i]$ and, relative to $V[y]$, measures the contribution of input x_i to the variance of the output y under the linear analysis model assumptions. The variance component $\beta_i^2 V[x_i]$ is usually estimated by way of a least squares estimate of β. We note in passing that interpretation of importance is not as straightforward when the inputs are not independent.

In uncertainty studies, the linear analysis model of eqn 9 is often used with only a subset x^s of size s of the input vector x appearing in the $x\beta$ term. The effects of the remaining components of x are collected in the error term e. This is the case in stepwise regression, for example, where a subset of the inputs is selected to be important and to form the regression model.

Let

$$x^s = (x_{i_1}, x_{i_2}, \cdots, x_{i_s}) \,, \tag{15}$$

be a vector of inputs chosen in a stepwise regression, say. Their subscripts are given by the set

$$I_s = \{i_1, i_2, \cdots, i_s\} \,, \tag{16}$$

and the corresponding vector of βs is

$$\beta^s = (\beta_{i_1}, \beta_{i_2}, \cdots, \beta_{i_s})^t \,. \tag{17}$$

The regression function $x^s \beta^s$ forms the basis of estimation and analysis, and assumes

$$\beta_i = 0 \text{ for } \; i \notin I_s \,. \tag{18}$$

Subsequent analyses assume that the subset regression model is a reasonable approximation to computation model and operate under the linear analysis assumptions

$$\begin{gathered} E(y \mid x^s) = x^s \beta^s \simeq m(x) \\ y = x^s \beta^s + e^s \\ E(e^s) = 0 \\ Cov[x^s \beta^s, e^s] = 0 \,. \end{gathered} \tag{19}$$

The vector β^s is estimated via the usual least squares as $\widehat{\beta^s}$, and the variance of $x^s\beta^s$ is estimated (with bias) with the estimator of β^s by

$$\widehat{V}[x^s\beta^s] = \widehat{\beta^s}^t V[x^s]\widehat{\beta^s}\,. \tag{20}$$

The regression also provides an estimate of the variance of e^s as the residual mean square. Therefore, the variance of y in eqn 14 for the subset regression model is estimated (with bias) by

$$\widehat{V}[y] = \widehat{\beta^s}^t V[x^s]\widehat{\beta^s} + \widehat{V}[e^s]\,. \tag{21}$$

When the inputs are independent, the estimated variance of y again can be written as

$$\widehat{V}[y] = \sum_{k\in I_s} \widehat{\beta_k^s}^2 V[x_k] + \widehat{V}[e^s]\,. \tag{22}$$

In this form, it is seen that the term involving the square of the estimator of β^s might be replaced by a bias-corrected estimator.

For a subset x^s of input variables, the measure of importance from the linear regression is a (multiple) correlation coefficient R^2 of the form

$$R^2 = \sum_{k\in I_s} \widehat{\beta_k^s}^2 V[x_k]/\widehat{V}[y]\,, \tag{23}$$

which is the fraction of the variance of y "explained" by the regression. This R^2 is similar to the usual one, except that the variance of x^s comes from its probability distribution rather than being estimated from a sample of values. In practice, though, the estimate of R^2 often will be the usual multiple correlation coefficient from sample values.

For subsets of size one of a single input variable, the quantity in eqn 23 is usually denoted by "ρ^2." For subsets of size greater than one, "R^2" is used. In this paper, the notations are used interchangeably to denote the correlation between a subset of inputs x^s, $s \geq 1$, and the model prediction y—with context indicating the size of the subset.

The derivations of this section show that the process of finding a good subset regression via stepwise regression, for example, and using the R^2 from the regression as an importance measure for the subset is really variance based, and that the methods are indeed suitable and proper as long as two assumptions are satisfied: (1) the linear analysis model is appropriate, and (2) the x^s is properly sampled from its own probability distribution.

The linear analysis model assumption combined with the additional assumption that the error variance, $V[e]$, is constant and independent of x is the basis for efficient estimation of the variance of the conditional expectation $x\beta$, meaning that many fewer computer runs are required than would be without the assumptions. The problem here, as expected, lies in the validity of the linear analysis model. The appropriateness of the linear analysis model is easily and properly questioned when considered against computation models that are known to be highly nonlinear in their inputs. A result of a breakdown in the linear analysis model assumption is unreliable estimation of the variance decomposition of eqn 12 and, thus, misleading indications or lack of indications of importances. The next section discusses methods[7] which do not rely on the linear analysis model. It shows how the methods compare to regression methods and how they might be used to complement regression methods.

4 NONPARAMETRIC VARIANCE-BASED METHODS

The nonparametric variance-based methods described in this paper assume that the input variables are partitioned in disjoint subsets $x = x^s \cup x^{(s)}$ such that the conditional variance $V[y \mid x^s]$ is not identically zero. The importance of the subset x^s is to be assessed. For an arbitrary subset x^s, the *general analysis model* is

$$\begin{aligned} y &= E(y \mid x^s) + e \\ E(e) &= 0 \\ Cov[E(y \mid x^s), e] &= \Sigma \,. \end{aligned} \tag{24}$$

No particular assumptions are made about the mathematical form of the conditional expectation $E(y \mid x^s)$ or the covariance matrix, Σ. In the linear analysis model of eqn 9, $E(y \mid x^s)$ is assumed to be of the form $x^s \beta^s$.

The basis of variance-based methods is the general analysis model in eqn 24 and the well-known variance decomposition,[8]

$$V[y] = V[E(y \mid x^s)] + E(V[y \mid x^s]) \,. \tag{25}$$

The decomposition is seen to be the same one used in linear regression methods in eqn 14 but without the assumed linear form $E(y \mid x) = x\beta$ of the conditional expectation of y. The objective of variance-based methods is to find subsets x^s of input variables for which the conditional expectation of y is a reasonable approximation to the computation model $m(\cdot)$. In other words, the objective, which parallels that of regression methods, is to find input subsets x^s for which

$$E(y \mid x^s) \simeq m(x) \tag{26}$$

for arbitrary form of $E(y \mid x^s)$.

The integral representation of eqn 25 is

$$\begin{aligned} &\int (y - \mu_y)^2 f_y(y) dy \;= \\ &\int (E(y \mid x^s) - \mu_y)^2 f_{x^s}(x^s) dx^s \;+\; \int\int (y - E(y \mid x^s))^2 f_{y|x^s}(y) f_{x^s}(x^s) dy dx^s \,. \end{aligned} \tag{27}$$

Examination of eqn 27 may make it clearer how the variance decomposition works. The variability in the random variable y is the left size of the equation,

$$V[y] = \int (y - \mu_y)^2 f_y(y) dy \,. \tag{28}$$

The amount of variability "explained" on the right side by input subset x^s is the variance of the expected value of y conditioned on x^s, namely,

$$V[E(y \mid x^s)] = \int (E(y \mid x^s) - \mu_y)^2 f_{x^s}(x^s) dx^s \,. \tag{29}$$

Estimation of eqn 25 parallels classical analysis-of-variance and is discussed in Section 6. The integral form in eqn 27 suggests that for estimation via the linear analysis model, the variance decomposition parallels the analysis of variance for regression analysis, namely,

$$\begin{gathered}\text{Total sum of squares } (\text{SST}) = \\ \text{Regression sum of squares } (\text{SSR}) + \text{ Error sum of squares } (\text{SSE})\,.\end{gathered} \tag{30}$$

The parallel is seen by the substitutions

$$\begin{gathered}\int \rightarrowtail \sum \\ \mu_y \rightarrowtail \overline{y} \\ E(y \mid x^s) \rightarrowtail x^s \widehat{\beta^s}\,.\end{gathered} \tag{31}$$

When the variables x^s are sampled appropriately, the three sums of squares, SST, SSR and SSE, can be used to estimate the terms in eqn 25.

Finally, the measure of importance of x^s with variance-based methods is the *correlation ratio*,

$$\eta^2 = \frac{V[E(y \mid x^s)]}{V[y]}\,. \tag{32}$$

As a measure influence, both the correlation ratio and its counterpart for the linear model, the correlation coefficient, were used by Pearson[9] at the turn of the twentieth century. A good development of both statistics appears in Kendall and Stuart.[10] Other references provide development and applications to computer models.[11,2,3,7]

The R^2 in eqn 23 from a regression model used to assess importance is really an estimator of the correlation ratio in eqn 32 under the assumption of the regression. That is, for the regression model $E(y \mid x^s) = x^s\beta^s$,

$$\begin{aligned}R^2 &\simeq \frac{(\beta^s)^t V[x^s]\beta^s}{V[y]} \\ &= \eta^2\,.\end{aligned} \tag{33}$$

When looking at a single input variable x, the correlation ratio from a linear regression model is equal to the square of the *correlation coefficient* ρ, defined by

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}\,. \tag{34}$$

From

$$\begin{aligned}\sigma_{xy} &= \text{Cov}(x, y) \\ &= \text{Cov}(x, E(y \mid x)) \\ &= \text{Cov}(x, x\beta) \\ &= \beta\sigma_x^2\,,\end{aligned} \tag{35}$$

the relationship between Equations 32 and 34 for a liner regression model is

$$\rho^2 = \eta^2 = \beta^2 \frac{\sigma_x^2}{\sigma_y^2}\,. \tag{36}$$

Therefore, for a single input variable and in stepwise linear regression, the R^2 of a model is an estimate of the correlation ratio of variance-based methods under a linear analysis model.

For the linear analysis model in the univariate case, we see that

$$\begin{aligned} \beta &= \frac{\sigma_{xy}}{\sigma_x^2} \\ &= \rho \frac{\sigma_y}{\sigma_x} \,. \end{aligned} \tag{37}$$

This extends to the multivariate case as

$$\begin{aligned} \beta &= V[x]^{-1} \sigma_{xy} \\ \beta^t V[x] \beta &= \sigma_{xy}^t V[x]^{-1} \sigma_{xy} \\ &= \eta^2 \times \sigma_y^2 \,. \end{aligned} \tag{38}$$

5 TRADE-OFFS

With linear regression methods, one may use relatively fewer computer runs to estimate variance components because of the assumed linear form for the conditional expectation of y. However, the relationship[10, p. 317]

$$\rho^2 \leq \eta^2 \tag{39}$$

shows that regression methods can miss important inputs that variance-based methods can find because ρ^2 can be small, indicating lack of importance, while η^2 can be large, properly indicating importance. In such cases, regression methods would fail to find important inputs because of a breakdown in assumptions of the linear analysis model. In practice, however, when estimates of the correlation ratio and the correlation coefficient are used simultaneously, the preference for the correlation ratio is not as unambiguous. That is, when the assumptions of the linear analysis model hold, meaning that $\eta^2 = \rho^2$, then the usual estimator of ρ^2 is more efficient than nonparametric estimate of η^2. In fact, the accuracy with which η^2 is estimated relative to that of ρ^2 can be disturbingly small. Therefore, it is very possible that $\widehat{\rho}^2$ could present a more accurate description of the importance of an input than would $\widehat{\eta}^2$, even in a nonlinear-model situation. Hence, it would be prudent to examine both estimators in analysis situations.

Under the linear analysis model in eqn 9, the form of the conditional expectation of y reduces the variance component estimation to the estimation of beta. For the general analysis model in eqn 24, the conditional expectation of y depends on x in an unspecified manner. Therefore, the conditional expectation and its variance are estimated via sampling theory. The number of computer runs necessary to assure adequate estimation is unknown in the general case because it will depend on the model $m(\cdot)$ under study. It seems reasonable, however, to assume that the number required is proportional to that required under a linear analysis model, and that the constant of proportionality depends on the complexity of $m(\cdot)$. The complexity of $m(\cdot)$ might be indicated by the number of parameters in a suitable Taylor series expansion of $m(\cdot)$. Therefore, variance-based methods may need many computer runs to properly identify important inputs.

The main point of this discussion is not whether ρ^2 (or η^2) works as a measure of importance, but that the source of breakdown of linear regression methods is in the linear analysis model assumption.

Therefore, the strength of variance-based methods is not so much in the use of variance ratios as it is in the use of a general analysis model. When the linear analysis model is valid, η^2 is estimated efficiently from a linear model. When the linear analysis model is not valid, nonparametric variance-based methods must be used.

6 DEMONSTRATION APPLICATION

In the following demonstration application, several points related to importance indicators and their estimation are discussed for a model with two input variables. The model output is a continuous function of the first input but a discontinuous function of the second. The response to the second input is intended to represent what might be encountered with a model that uses (discontinuous) categorical variables as inputs. Results from a simulation study are presented as examples of what one might encounter in practice.

6.1 Importance measures

As shown earlier, regression methods are a subset of nonparametric variance-based methods. In variance-based methods, the importance of an input x is measured by the correlation ratio from eqn 32. Regression methods assume that the functional form of the conditional expectation of y, which appears in the numerator of the expression for η^2, is known. Under the assumption that $E(y \mid x) = x\beta$, the correlation ratio is equal to the square of the (multiple) correlation coefficient ($R^2 = \rho^2$) in eqn 33. For this example, we use ρ^2 to denote either the ordinary correlation coefficient for a single x, or the multiple correlation coefficient under a linear analysis model for multivariate x.

6.2 Estimation of importance measures

The correlation ratio and the correlation coefficient are estimated (with bias) from a sample of values as follows. Let the values

$$\{x_i^s,\ i = 1, \cdots, n\} \tag{40}$$

of the input subset of size $s = 1$ be a random sample of size n from the input distribution f_{x^s}.. Let

$$\{y_{ik},\ k = 1, \cdots, r\} \tag{41}$$

be a conditionally independent random sample of size r from $f_{y|x_i^s}$ for $i = 1, \cdots, n$. The r replicates are generated by sampling the other inputs x not in x^s. Let the sample means be

$$\overline{y}_i = \frac{1}{r}\sum_{k=1}^{r} y_{ik}\ , \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} \overline{y}_i\ , \quad \text{and } \overline{x}^s = \frac{1}{n}\sum_{i=1}^{n} x_i^s\ . \tag{42}$$

A decomposition of sums of squares paralleling an analysis of variance is given in Table 1. Estimates of the correlation ratio η^2 and the square of the correlation coefficient ρ^2 are indicated at the bottom of the table. An advantage of using the estimators indicated is that they have the property of their population counterparts that $0 \le \widehat{\rho}^2 \le \widehat{\eta}^2 \le 1$, as implied in the table.

The expected value of the total sum of squares SST in Table 1 is

$$\begin{aligned} E(\text{SST}) &= r(n-1)V[y] + (r-1)E(V[y \mid x^s]) \\ &= r(n-1)\sigma_y^2 + (r-1)\overline{\sigma}_e^2 \,, \end{aligned} \tag{43}$$

where $\overline{\sigma}_e^2 = E(V[y \mid x^s])$ is an error term corresponding to the variability $\sigma_y^2 = V[y]$ in y not attributable to x^s. The total mean square, $\text{SST}/(nr-1)$, can be driven to σ_y^2 with n for fixed r. Similarly,

$$\begin{aligned} E(\text{SSB}) &= r(n-1)V[E(y \mid x^s)] + (n-1)E(V[y \mid x^s]) \\ &= r(n-1)V[E(y \mid x^s)] + (n-1)\overline{\sigma}_e^2 \,, \end{aligned} \tag{44}$$

and

$$\begin{aligned} E(\text{SSW}) &= n(r-1)E(V[y \mid x^s]) \\ &= n(r-1)\overline{\sigma}_e^2 \,. \end{aligned} \tag{45}$$

With a caution about biased estimation in mind, we proceed to describe the demonstration model.

Table 1. **Analysis-of-variance decomposition of sums of squares**

Source of Variation	Degrees of Freedom	Sum of Squares
Total	$nr-1$	$\text{SST} = \sum_{i=1}^{n}\sum_{k=1}^{r}(y_{ik}-\overline{y})^2$
Between	$n-1$	$\text{SSB} = r\sum_{i=1}^{n}(\overline{y}_i-\overline{y})^2$
Regression	1	$\text{SSR} = r\left[\sum_{i=1}^{n}(\overline{y}_i-\overline{y})(x_i-\overline{x}^s)\right]^2 / \sum_{i=1}^{n}(x_i^s-\overline{x}^s)^2$
Error (lack of fit)	$n-2$	$\text{SSE} = \text{SSB} - \text{SSR}$
Within	$n(r-1)$	$\text{SSW} = \sum_{i=1}^{n}\sum_{k=1}^{r}(y_{ik}-\overline{y}_i)^2$

$$\begin{aligned} \widehat{\eta}^2 &= \text{SSB}/\text{SST} \\ \widehat{\rho}^2 &= \text{SSR}/\text{SST} \end{aligned}$$

6.3 Demonstration model

The demonstration application is designed to show the range of results of analyses that can be encountered for continuous inputs and categorical inputs and for regression and nonparametric variance-based methods. (Consideration of alternative measures of association for categorical variables are not included in this paper.) In the notation of the previous sections, the bivariate input vector x is $x = (t, d)$, and subsets of the input vector are $x^s = t$ or $x^s = d$. The model prediction y, given by $m(\cdot)$, is a randomly selected polynomial

$$y = m(t, d) = \text{Legendre polynomial in } t \text{ of degree } d\,, \tag{46}$$

with inputs t and d for which

$$\begin{aligned} t \sim\ & \text{Uniform on } [-1, +1] \text{ is the variable in the polynomial,} \\ d \sim\ & \text{Uniform on } \{1, 2, 3, 4, 5\} \text{ is the degree of the polynomial, and} \\ & t \text{ and } d \text{ are statistically independent.} \end{aligned} \tag{47}$$

For example, with probability $1/5$, the model prediction y is a Legendre polynomial in t of degree $d = 3$. The different polynomials represent 5 different responses of y to t. The responses of y as a function of t for the 5 polynomials are plotted in Fig. 1. Legendre polynomials have the properties that they are orthogonal and integrate to 0 on the interval $[-1, 1]$. The mean value and variance of y are

$$\begin{aligned} E(y) &= 0 \\ V[y] &= 3043/17325 \simeq 0.1756\,. \end{aligned} \tag{48}$$

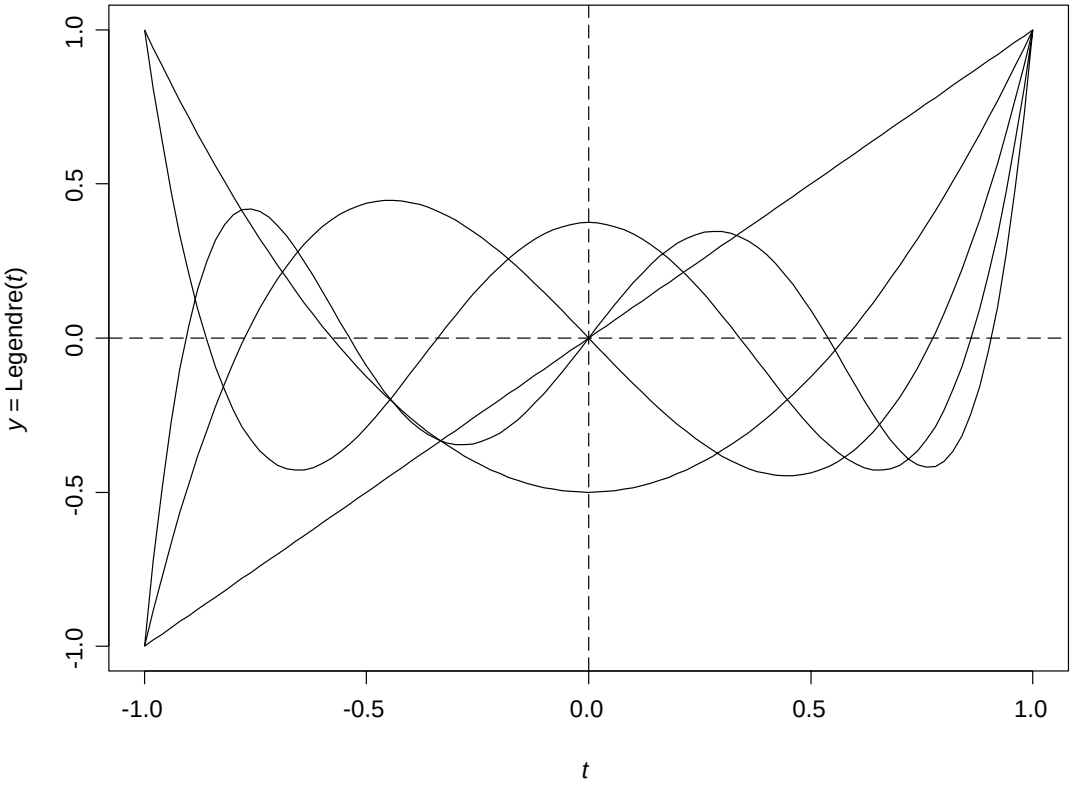

Fig. 1. Legendre polynomials of degree d from 1 to 5.

6.4 Importance of inputs

The model prediction y depends on two input variables. Fig. 1 illustrates "importance" of t by the change in y while moving along any one of the 5 curves. The importance of t changes with d. Similarly, "importance" of d is indicated by the change in y as one moves among the 5 curves in a vertical direction for fixed t. The importance of d changes with t. Local importance relates to the change in mean of y conditioned on t or d. Conditioned on $d \in \{1, 2, 3, 4, 5\}$, the mean and variance of y are given by

$$\begin{aligned} E(y \mid d) &= \frac{1}{2}\int_{-1}^{+1} \text{Legendre}(t, d)dt = 0 \\ V[y \mid d] &= \frac{1}{2}\int_{-1}^{+1} (\text{Legendre}(t, d) - E(y \mid d))^2 dt = \frac{1}{2d+1}\,. \end{aligned} \tag{49}$$

Conditioned on t, the mean $E(y \mid t)$ and variance $V[y \mid t]$ of y are algebraically more complicated than those for d in eqn 49. They are plotted in Fig. 2. The mean is a smoothly varying, generally increasing function of t. The variance is maximum at $t = -1$ and is 0 at $t = +1$. The variance is roughly constant between –0.5 and 0.5.

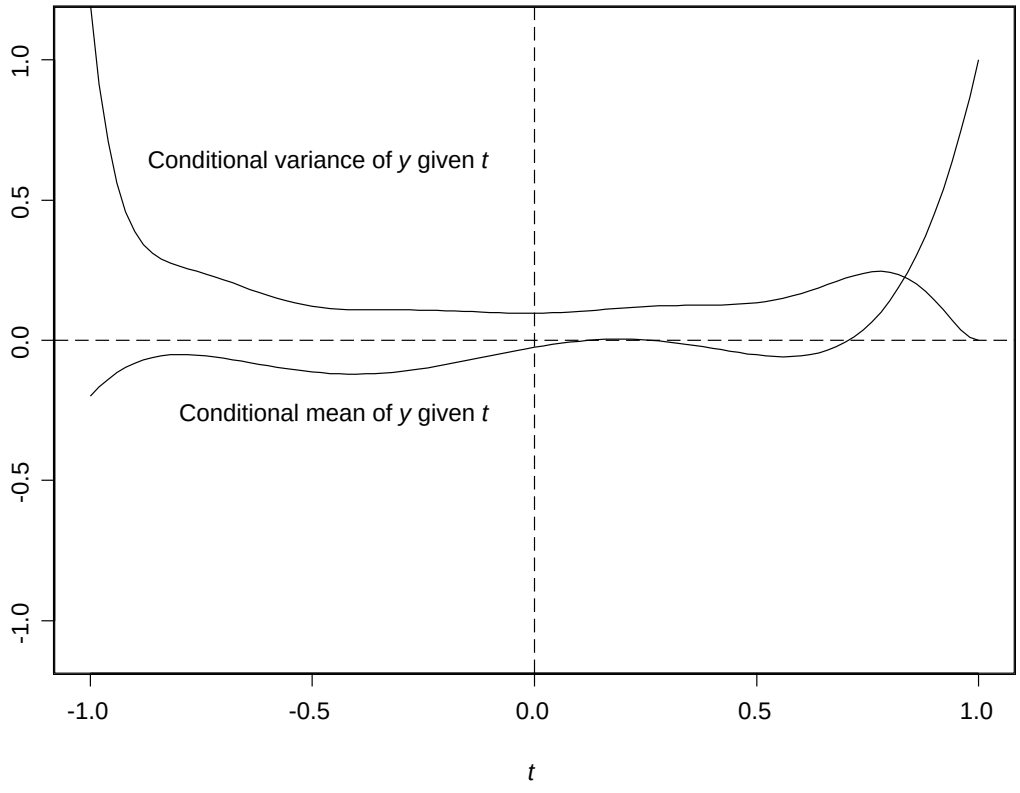

Fig. 2. Conditional mean $E(y \mid t)$ and variance $V[y \mid t]$.

Importance measures, whether they are η^2 or ρ^2, are intended to show how an input influences the value of an output. In variance-based methods, we look to see how an input controls the value of

the output in the sense of how much the variance of the output is reduced when the value of the input is held fixed. Specifically, η^2 measures how much, on average, the variance of y conditioned on an input is reduced relative to the unconditional variance. Equivalently, η^2 measures how closely the variance of the conditional expected value of y matches, on average, the unconditional variance of y.

Variance of the conditional expectation may not always be an appropriate measure of importance, however. For example, we see from Equation 49 that, for all values of d, the expected value of y given d is a constant value of 0. Therefore, the variance of the conditional expected value of y is 0, and d is measured to be unimportant by variance-based methods, in general, including regression methods, even though the value of y varies with d in a prescribed manner.

Variance of the conditional expectation for the input t behaves more as anticipated. True population (theoretical) values of η^2 from eqn 32 and of ρ^2 from eqn 34 are given in Table 2. Values of η^2 are computed from the variance components given in Table 3. Those values were calculated for t and d from quantities like those in eqn 49 for d. The first row of numerical values in Table 2 corresponds to Fig. 1, where y is a random polynomial of degree up to $d_{max} = 5$. For comparison, alternative models which limit the degree of y from 1 to 5 and to 10 are given in the last 6 rows of the table. The table shows that for low degrees, both η^2 and ρ^2 behave as anticipated for t, but that in the limit, both indicators go to zero.

The second column of Table 3 shows that the expected value of y given d does not change. Therefore, both the correlation ratio η^2 and the correlation coefficient ρ^2 for d are 0 in Table 2, indicating (erroneously) not only that d is unimportant, but that it is a completely irrelevant input with respect to the importance measures. However, while it is true that the expected value of y given d is constant for all d, the variance of y given d is not, as shown in Equation 49. Therefore, one is led to the observation that variance-based measures of importance, which use the variance of the conditional expectation of y, may not always be effective as measures of importance. Thus, notions of importance in addition to the variance of the conditional expectation may be necessary for assessing uncertainty importance in some models.

6.5 Estimators $\widehat{\eta}^2$ and $\widehat{\rho}^2$ and their sampling variability

Sampling variability of estimators of the correlation ratio and the correlation coefficient for t and d can be significant. To get an idea of how big the variability can be, estimators of the importance measures were calculated from 100 independent sets of samples of various sizes N. $N = n \times r$ is the product of a number of distinct values (n) and the number each is replicated (r), which are different for t and d. For each sample set of size N computer runs, there are $n = n_t$ distinct values of the continuous input t replicated $r = r_t$ times, and $n = n_d$ distinct values of the discrete input d replicated $r = r_d$ times. The numbers n_t and r_t were integers chosen so that

$$n_t \times r_t = N \; . \tag{50}$$

The input d has only $d_{max} = 5$ distinct values. The numbers n_d and r_d satisfy

$$\begin{aligned} n_d &= 5 \\ r_d &= N/5 \; , \end{aligned} \tag{51}$$

and were made integers by choice of N.

Table 2. **Importance measures with** $y =$ **Legendre**(t, d) **for** d **uniform on** $1, 2, \cdots, d_{max}$ **and** t **uniform on** $[-1, 1]$

d_{max}	ρ_d^2	η_d^2	ρ_t^2	η_t^2
Values for model used in demonstration application				
5	0	0	.08	.20
For comparison, values for polynomial models with other maximum degree d_{max}				
1	—	—	1.0	1.0
2	0	0	.31	.50
3	0	0	.16	.33
4	0	0	.11	.25
5	0	0	.08	.20
10	0	0	.03	.10

Table 3. **Components of** η^2 **for** $d_{max} = 5$

Input	$E(y \mid \text{input})$	$V[E(y \mid \text{input})]$	$E(V[y \mid \text{input}])$	$V[y]$
d	0	0	$\frac{3043}{17325} \simeq 0.1756$	$\frac{3043}{17325} \simeq 0.1756$
t	polynmial in t (See Figure 2)	$\frac{3043}{86625} \simeq .0351$	$\frac{12172}{86625} \simeq 0.1405$	$\frac{3043}{17325} \simeq 0.1756$

The minimum number of replicated values required for estimation is $r = 2$. Figs. 3, 4, and 5 show histograms of values of estimators of the correlation ratio and the square of the correlation coefficient (both defined in Table 1) for increasing total sample sizes $N = 10,\ 20,$ and 100, with $r_t = 2$ and $r_d = N/5$.

The upper left graph in each of the figures is a histogram of values of the estimator $\widehat{\eta}_t^2$ of the correlation ratio for t. They have vary large spreads for $N = 10$ and 20 in Figs. 3 and 4, and converge to a biased estimate as indicated by their distribution for $N = 100$ in Fig. 5. To get an

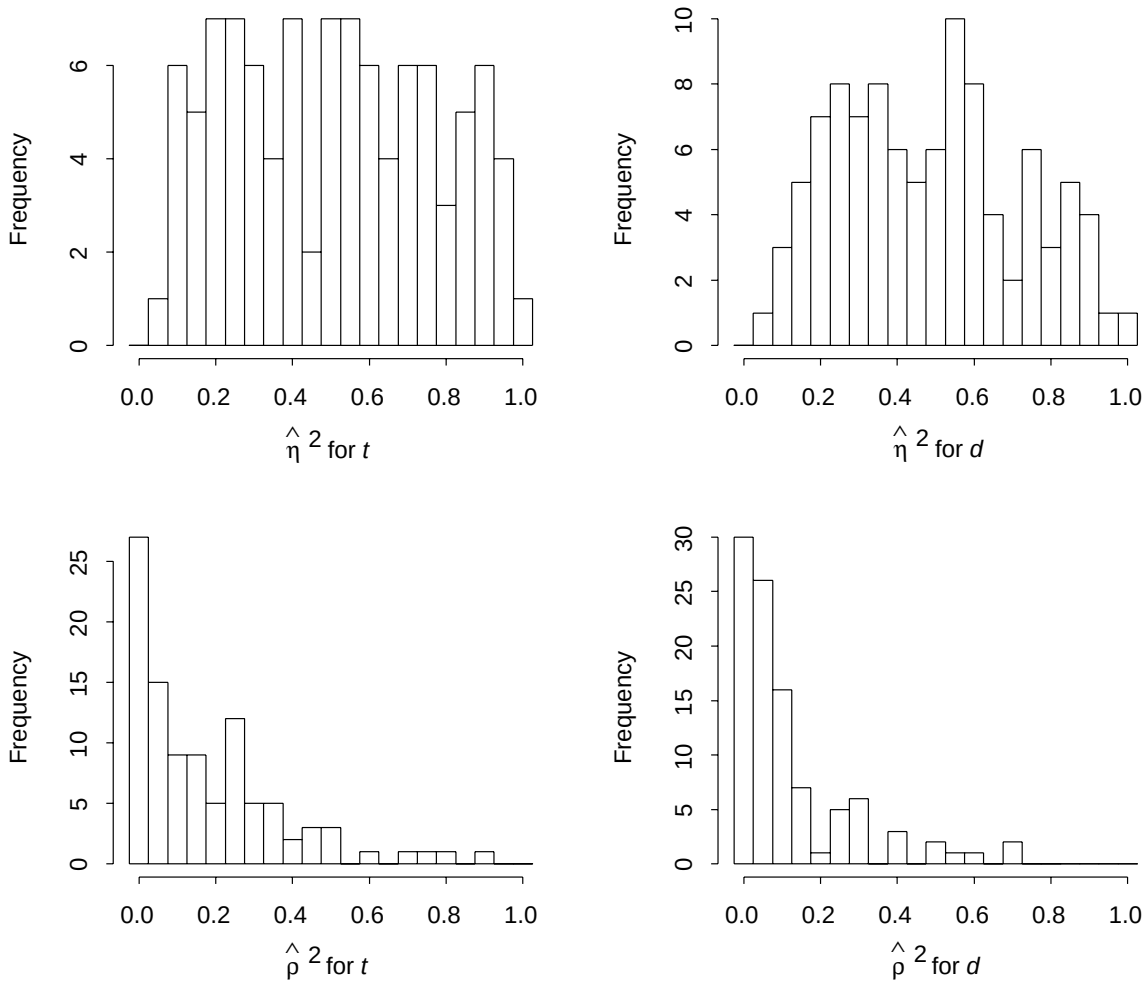

Fig. 3. $N = 10,\ r_t = 2,\ r_d = 2$ in 100 simulations.

idea of the bias, one can examine the expected values of the sums of squares in the estimator of η^2. The ratio of expectations is given by

$$\frac{E(\text{SSB})}{E(\text{SST})} = \frac{(n-1)(r-1)\eta^2 + (n-1)}{rn - 1 - (r-1)\eta^2}\,. \tag{52}$$

In the limit with n,

$$\lim_{n\to\infty} \frac{E(\text{SSB})}{E(\text{SST})} = \eta^2 + \frac{1}{r}(1-\eta^2) \tag{53}$$

While the ratio of expectations is not equal to the expectation of the ratio, the result shows the order of the bias with r in the estimator $\widehat{\eta}_t^2$. In this demonstration, for which $\eta_t^2 = 0.2$, the limiting value of the ratio of expectations is 0.6, clearly consistent with Fig. 5. Fig. 6 indicates more succinctly for $r_t = 2$ the convergence of the biased estimator with n_t. To complete the picture, Fig. 7 shows the convergence with r_t of the estimator to the population value of 0.2 for $n_t = 50$.

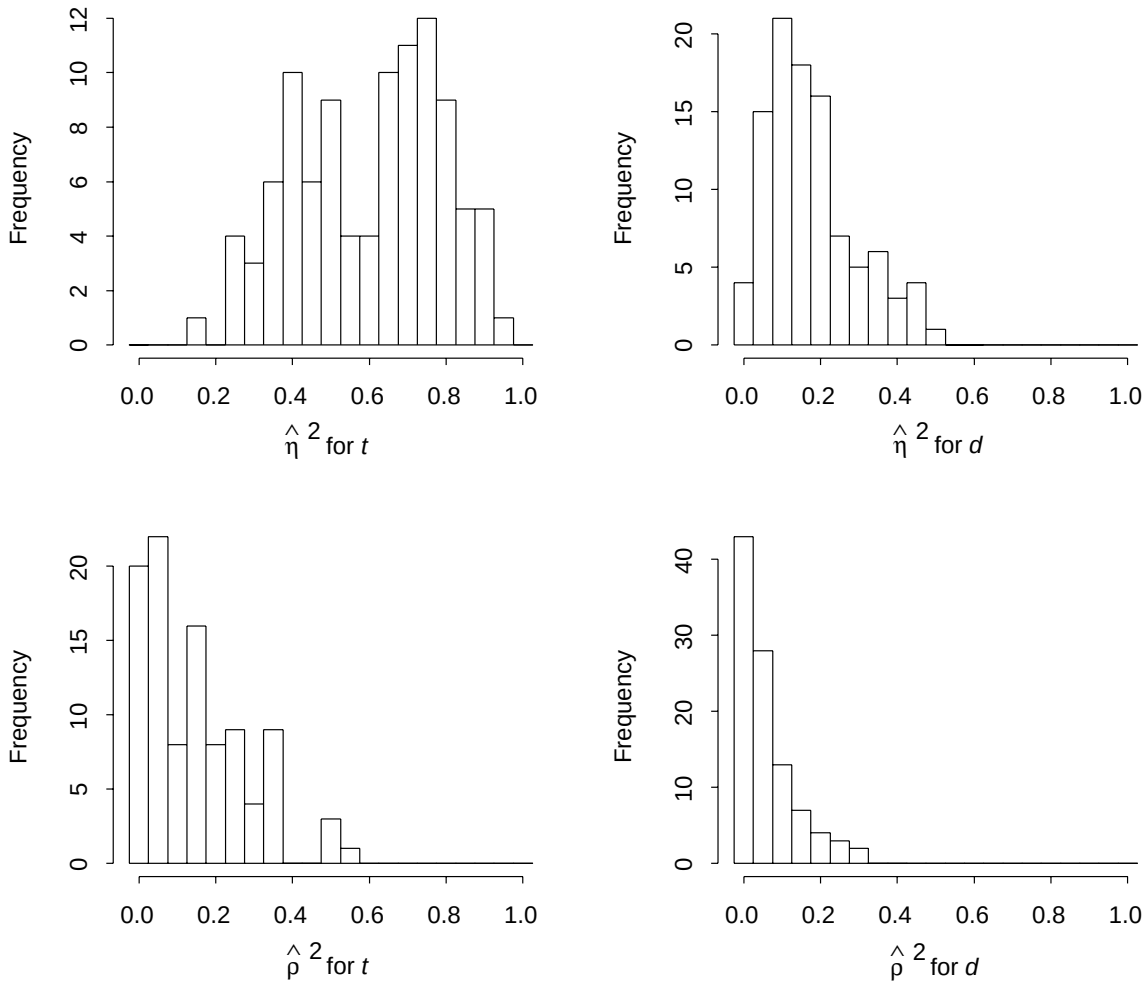

Fig. 4. $N = 20,\ r_t = 2,\ r_d = 4$ in 100 simulations.

The upper right graph in each of the Figs. 3, 4, and 5 is a histogram of the estimator of the correlation ratio for d. The population value of the correlation ratio for d is 0. Therefore, we compare the simulation results with the limit with n for $\eta^2 = 0$ in

$$\lim_{n\to\infty} \frac{E(\text{SSB})}{E(\text{SST})} = \lim_{n\to\infty} \frac{n-1}{rn-1} = \frac{1}{r}\,. \tag{54}$$

The theoretical result is consistent with the results of Fig. 5, even though d has only 5 distinct values.

The lower left and right graphs in the figures pertain to estimators of the square of the correlation coefficient. Results for $N = 10$ in Fig. 3 indicate, as in the case of the correlation ratio, that the sample size is too small for meaningful conclusions. Fig. 5 results are consistent with theoretical results for the correlation coefficient, that say (erroneously) neither t nor d is an important input.

We would conclude from Fig. 5 that (1) as measured by ρ^2, both t and d are unimportant, and (2) as measured by η^2, t is important but d is unimportant. These conclusions are consistent with theory and point out inadequacy in ρ^2 when the linear analysis model is not valid, and the weakness of using only the variance of the conditional expectation as a measure of importance under the general analysis model.

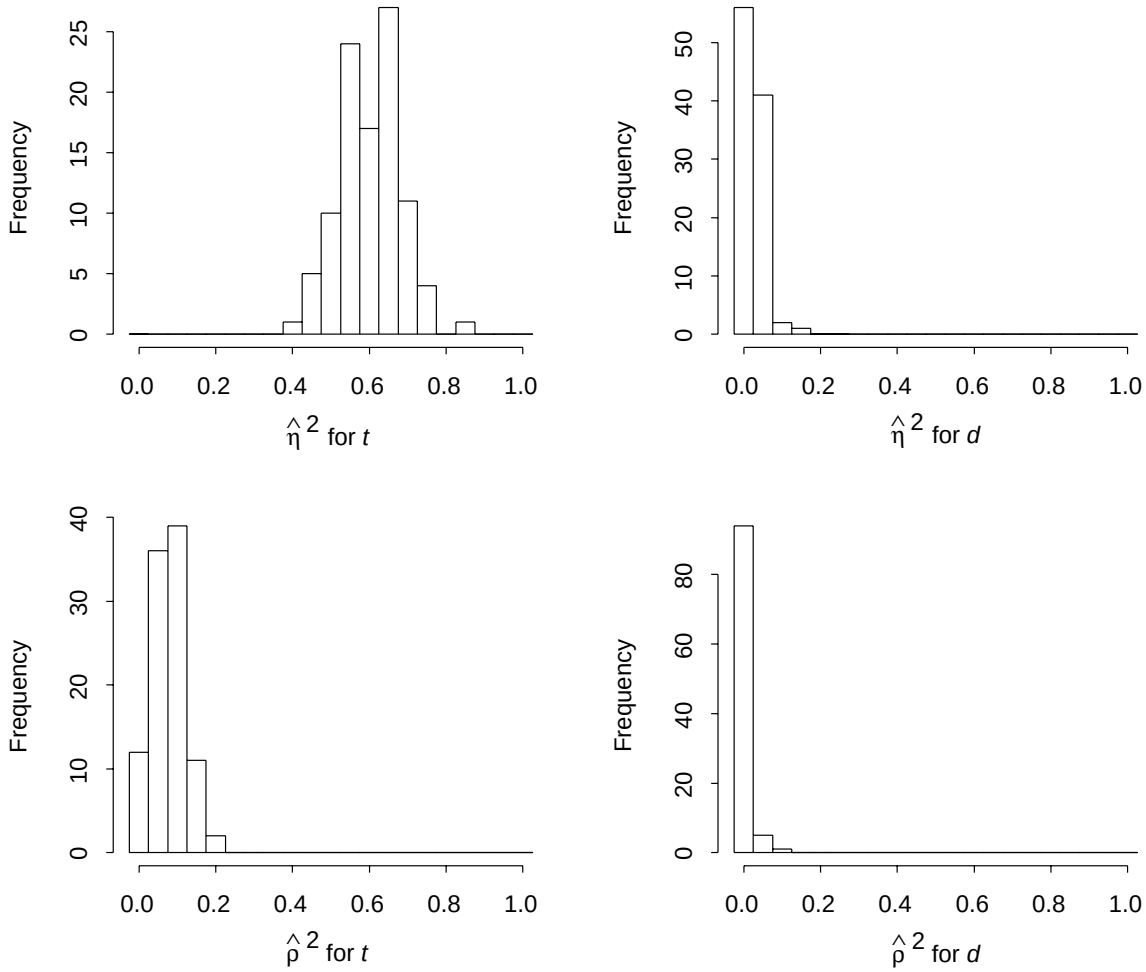

Fig. 5. $N = 100,\ r_t = 2,\ r_d = 20$ in 100 simulations.

The widths of the histograms in Figs. 3, 4, and 5 point out that for the same sample size, N, the variability in $\widehat{\eta}^2$ is greater than that in $\widehat{\rho}^2$. Therefore, larger sample sizes are required for estimates of the correlation ratio than for estimates of the correlation coefficient for specified degree of precision.

Examples of bias and precision of $\widehat{\eta}^2$ are shown in Figs. 6 and 7. For example, $\widehat{\eta}_t^2$ converges with n to a somewhat biased value using only $r_t = 2$ replicates, as indicated in Fig. 6 for increasing sample sizes $n = 5,\ 10,\ 50,\ 100,\ 1000$. Sampling variability of $\widehat{\eta}^2$ is indicated by the widths of the box plots. Fig. 7 illustrates similar convergence with r for sample size $n = 50$. As before, sampling variability of $\widehat{\eta}^2$ is indicated by the size of the box plots.

For comparison between the correlation ratio and the correlation coefficient, the effects of sample sizes of distinct values n and number of replicates r are summarized in Figs. 8 and 9. Figure 8 illustrates by the decreasing sizes of the box plots that a substantial number of computer runs may be necessary to adequately estimate the correlation ratio. Figure 9 illustrates by the approaching of centers of the box plots to zero that the correlation coefficient can fail to detect important inputs.

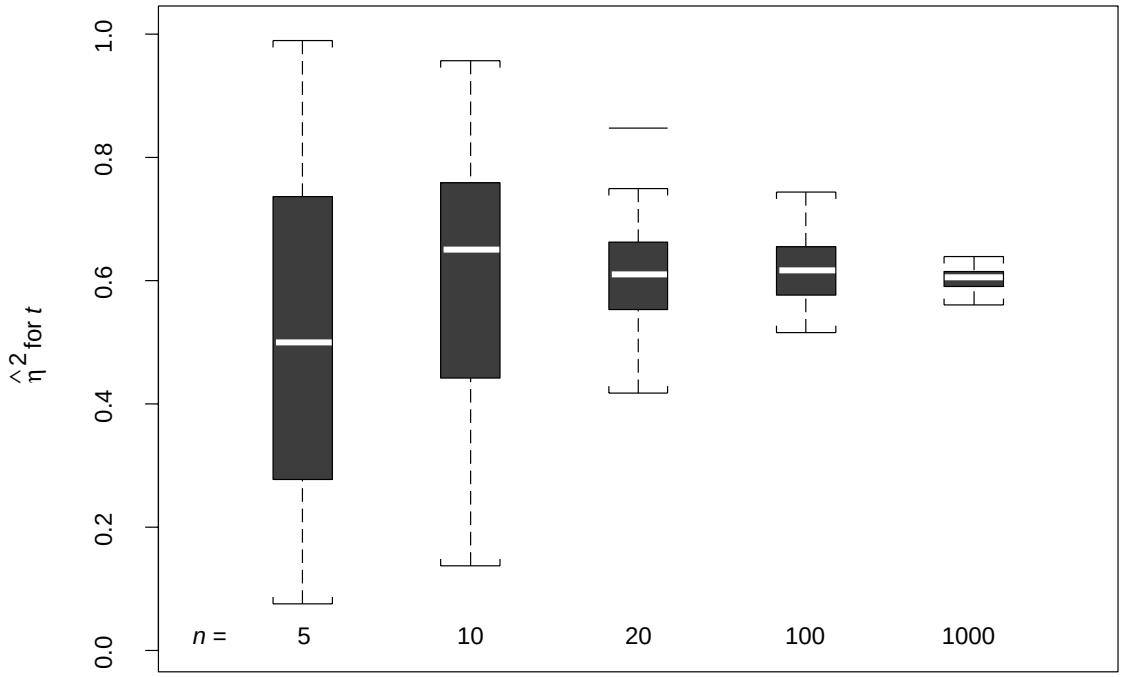

Fig. 6. Convergence of the estimated correlation ratio for t to a biased value with $n = 5,\ 10,\ 50,\ 100,\ 1000$ for $r = 2$, where each box plot contains 100 simulations.

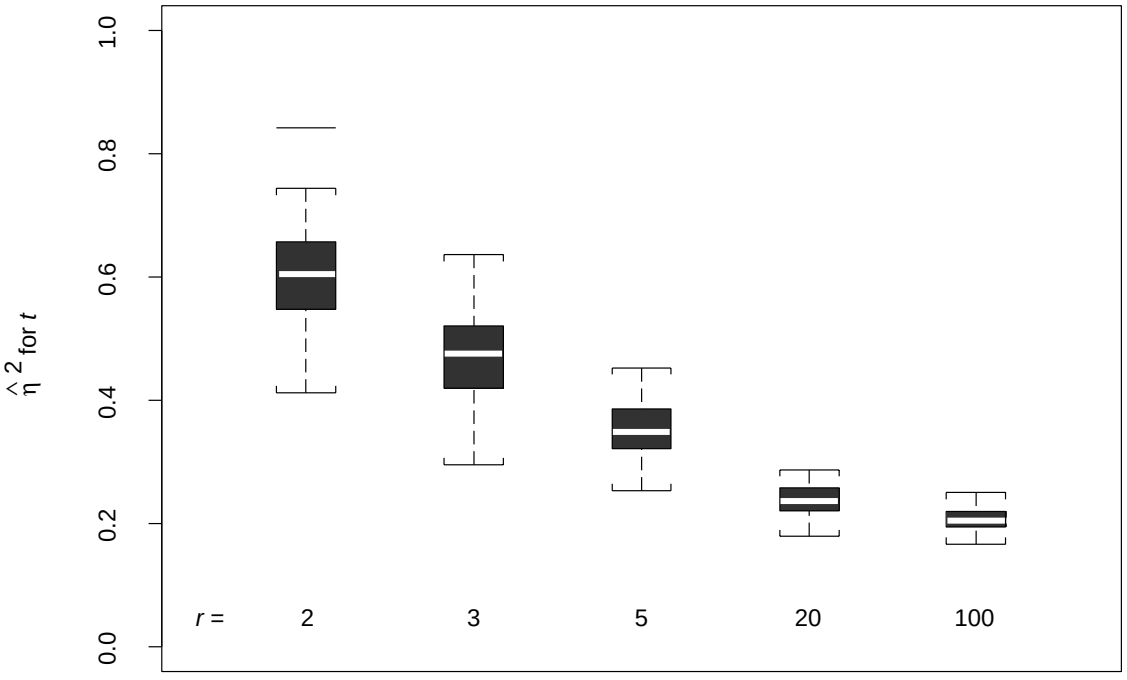

Fig. 7. Convergence of the estimated correlation ratio for t with $r = 2,\ 3,\ 5,\ 20,\ 100$ for $n = 50$, where each box plot contains 100 simulations.

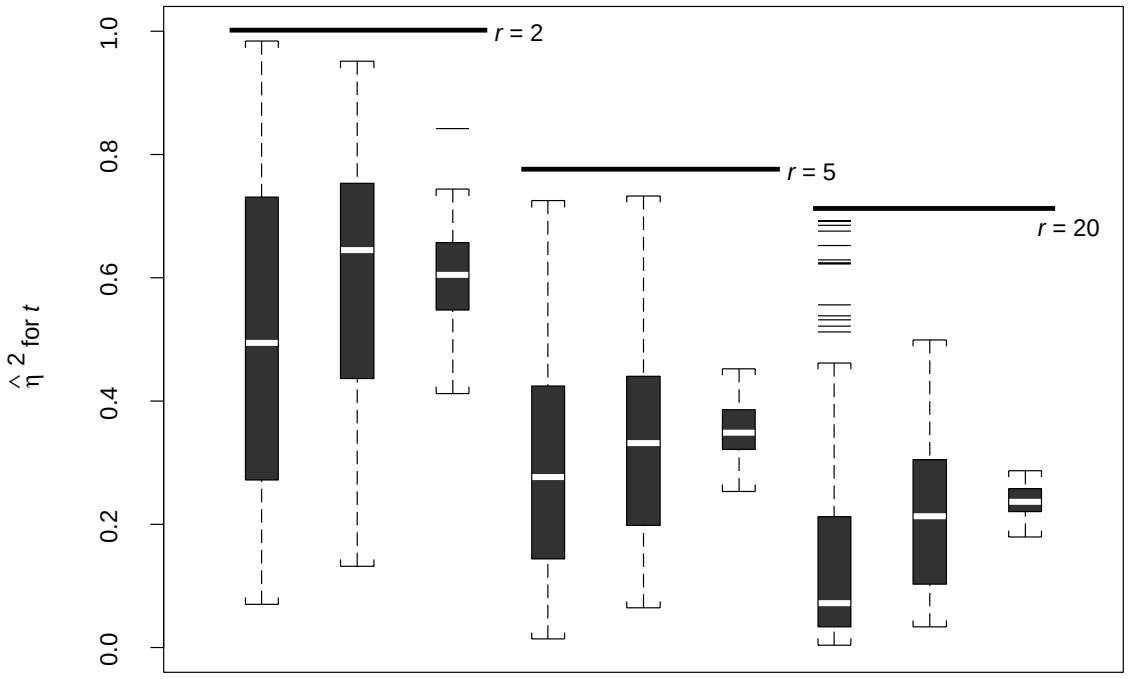

Fig. 8. Estimates of the correlation ratio for t with $n = 5,\ 10,\ 50$ grouped by $r = 2,\ 5,\ 20$.

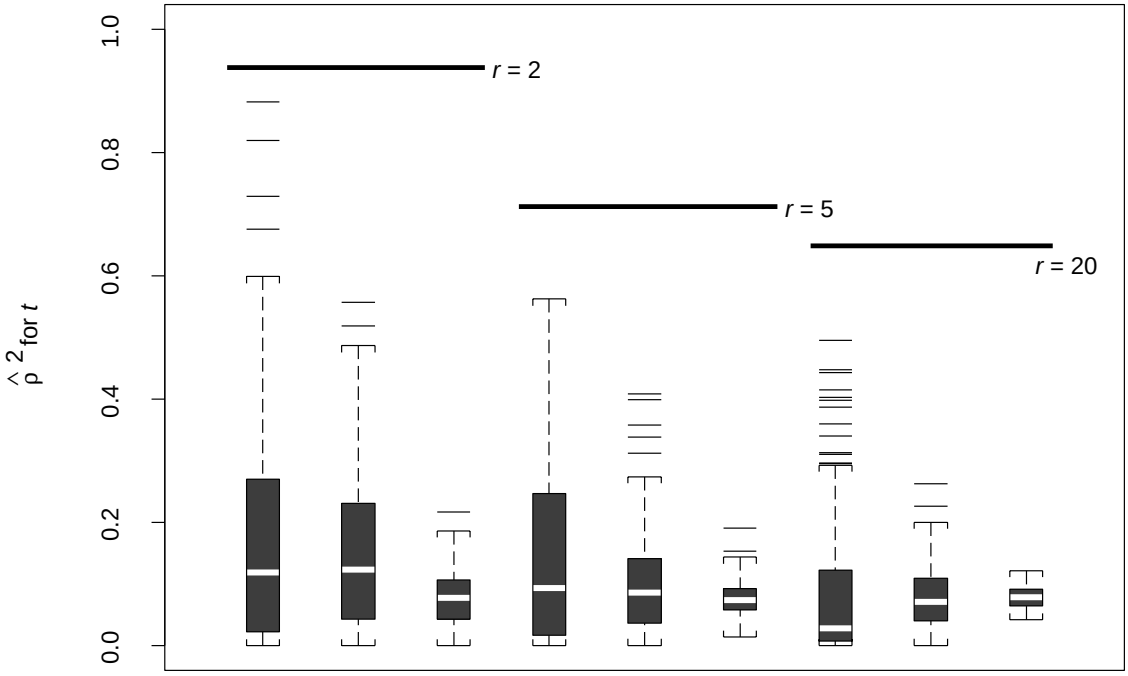

Fig. 9. Estimates of the square of the correlation coefficient for t with $n = 5,\ 10,\ 50$ grouped by $r = 2,\ 5,\ 20$.

7 CONCLUSIONS AND RECOMMENDATIONS

Nonparametric variance-based methods are theoretically superior to linear regression methods, which are expected to fail to detect important inputs in some situations when models are nonlinear. However, nonparametric variance-based methods can require substantially more computer runs to be effective. Therefore, this paper recommends that the nonparametric variance-based correlation ratio and the (multiple) correlation coefficient from linear regression both be used in analysis of uncertainty of model predictions. Validity of assumptions of the analysis model for regression methods and sufficiency of the sampling design, particularly for nonparametric variance-based methods, are so important that techniques for evaluating their adequacy ought to be a regular part of analysis procedures.

8 ACKNOWLEDGMENTS

This work was supported by the United States Nuclear Regulatory Commission, Office of Nuclear Regulatory Research, under Job Code W6505. The author is grateful for many valuable conversations with Richard J. Beckman of the Los Alamos National Laboratory.

REFERENCES

1. NUREG–1150. Severe accident risks: An assessment for five U.S. nuclear power plants. Tech. Rep. NUREG–1150, U.S. Nuclear Regulatory Commission, 1990.
2. Iman, R. L. & Hora, S. C. A robust measure of uncertainty importance for use in fault tree system analysis. *Risk Analysis 10*, 3 (1990), 401–406.
3. Saltelli, A., Andres, T. H., & Homma, T. Sensitivity analysis of model output: An investigation of new techniques. *Computational Statistics & Data Analysis 15* (1993), 211–238.
4. Homma, T. & Saltelli, A. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering and System Safety 52* (1996), 1–17.
5. Sobol', I. M. Sesitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiment 1* (1993), 407–414.
6. Stein, M. Large sample properties of simulations using Latin hypercube sampling. *Technometrics 29*, 2 (1987), 143–151.
7. McKay, M. D. Evaluating prediction uncertainty. Tech. Rep. NUREG/CR–6311, U.S. Nuclear Regulatory Commission and Los Alamos National Laboratory, 1995.
8. Parzen, E. *Stochastic Processes*, page 55. Holden Day, San Francisco, 1962.
9. Pearson, K. Mathematical contributions to the theory of evolution. *Proceedings of the Royal Society of London 71* (1903), 288–313.
10. Kendall, M. & Stuart, A. *The Advanced Theory of Statistics*, fourth ed., vol. 2. MacMillan Publishing Co., New York, 1979, ch. 26.
11. Krzykacz, B. Samos: A computer program for the derivation of empirical sensitivity measures of results from large computer models. Tech. Rep. GRS-A-1700, Gesellschaft fur Reaktorsicherheit (GRS) mbH, Garching, Republic of Germany, 1990.